Stability selection:
enhanced variable selection and network models

Session 3/3

Barbara Bodinier

Computational Epidemiology
MSc Health Data Analytics and Machine Learning

4 March 2021

Imperial College
London

Back to the multi-OMICs signature of smoking

- Objective: characterisation of the molecular signature of tobacco smoking
- In particular, exploring how CpG sites and transcripts jointly mediate the effect of smoking

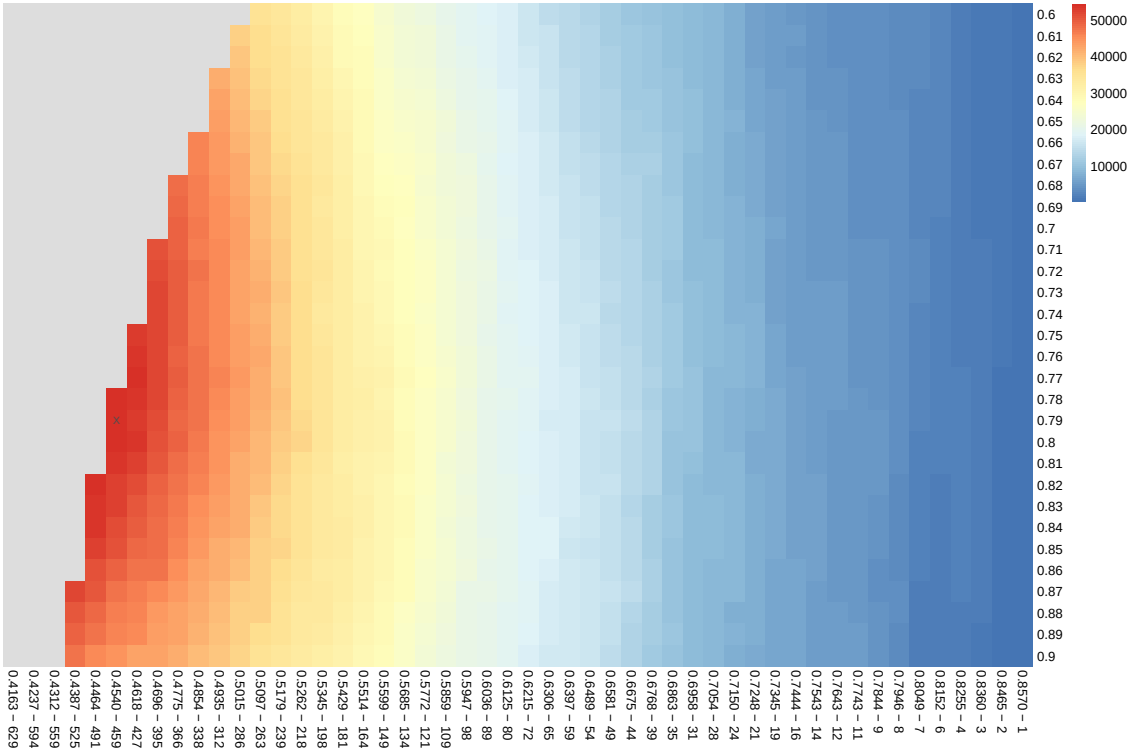
- Estimation of pairwise relationships in conditional independence networks

1. Methylation networks: epigenetic response to smoking
   $\Rightarrow$ Need for a metric measuring the correlation between variables
   $\Rightarrow$ Calibration issue: how many edges to include?

2. Multi-OMICs networks: integration of data from multiple OMICs platforms for a better understanding of the molecular consequences of smoking
   $\Rightarrow$ Accommodate heterogeneous blocks of data

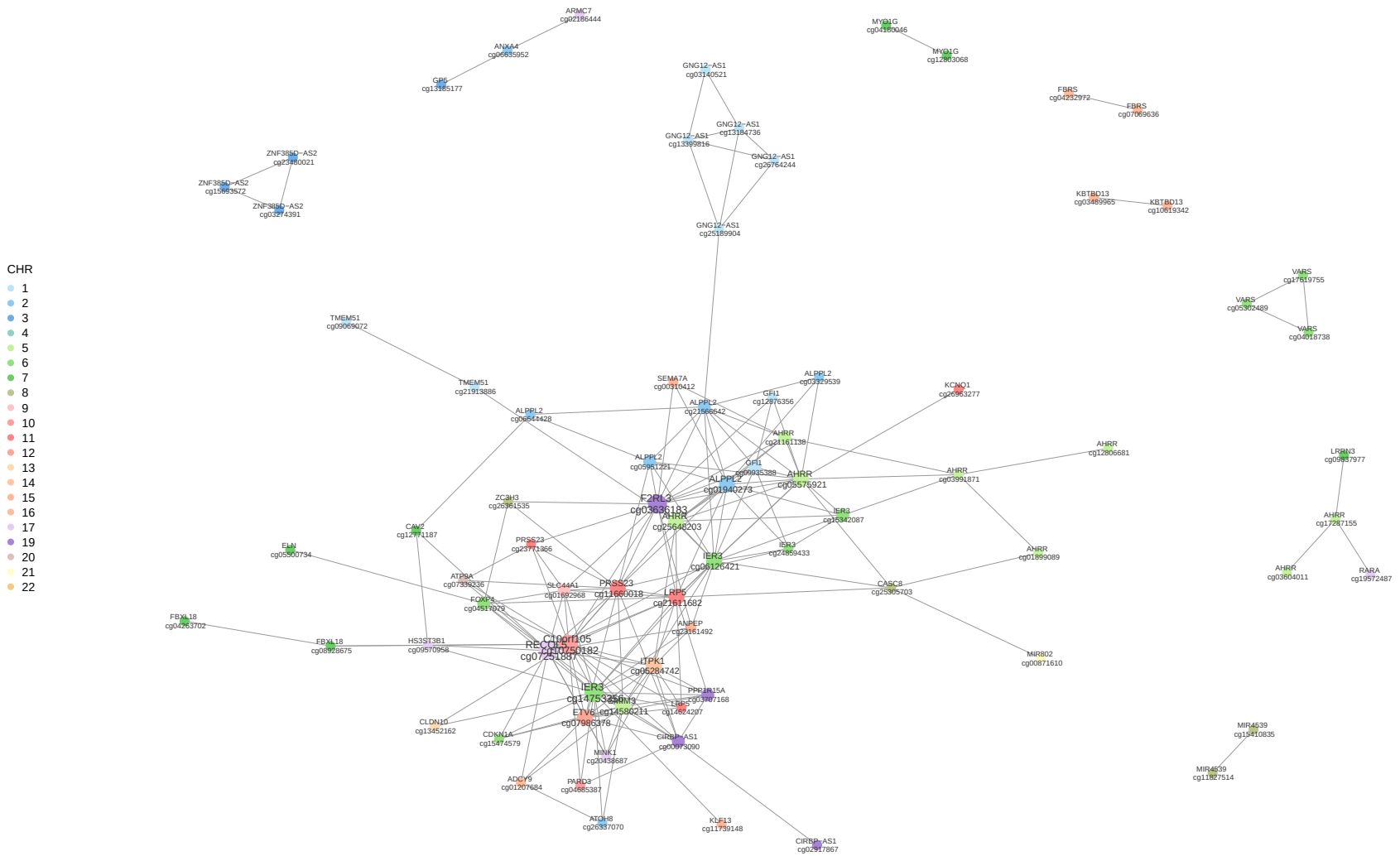
Network of smoking-related CpG sites

- Focus on 159 DNA methylation markers of long-term exposure to tobacco smoking (meta-analysis by London et al.)
- Measurements from 250 Women from the NOWAC cohort
- Calibration maximising the stability score under constraint that PFER < 30



$\Rightarrow$ Calibrated λ = 04540 and π = 0.79

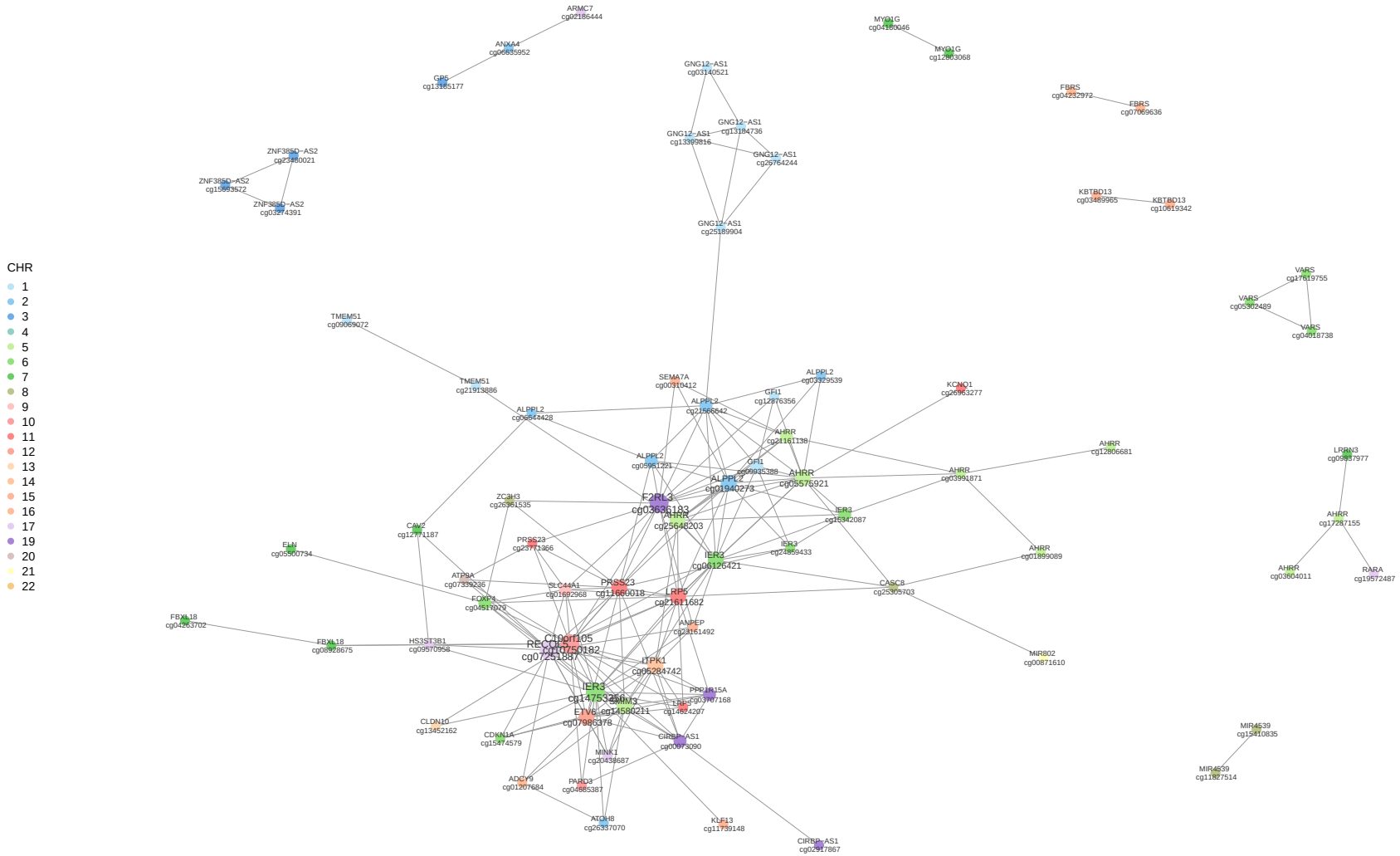
Network of smoking-related CpG sites



$\Rightarrow$ Modules of CpG sites from the same chromosome/closely located on the genome are detected

$\Rightarrow$ Both cys- and trans-relationships are detected

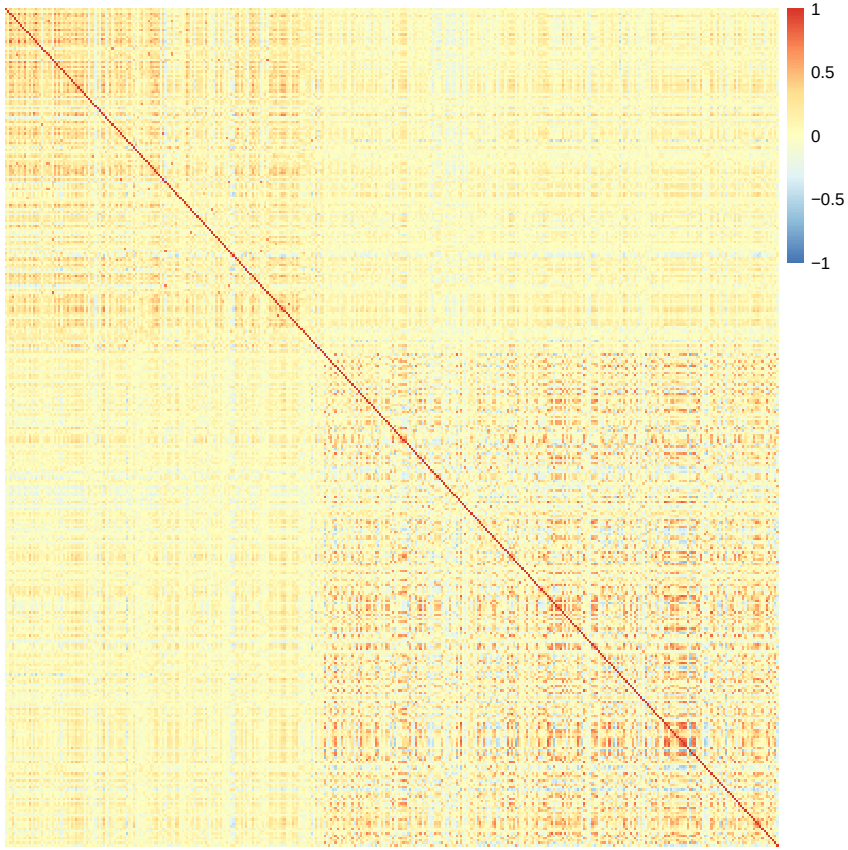
Network of smoking-related CpG sites



⇒ Central role of F2RL3, AHRR and ALPPL2 (high degree)

Towards OMICs integration

- Pearson's correlation heatmap between CpG sites and transcripts (left)

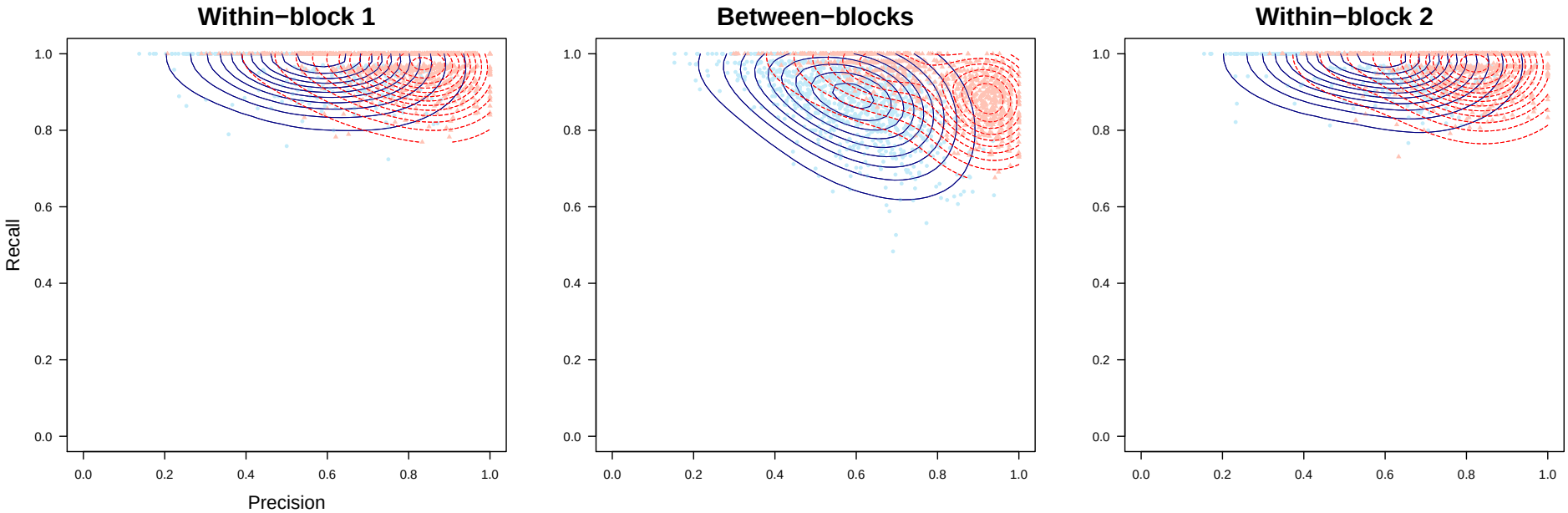


$\Rightarrow$ Overall weaker between-OMICs than within-OMIC correlations

$\Rightarrow$ Multi-block calibration with block-specific parameters to account for heterogeneity in the data

Performance of multi-block calibration

- Block-specific precision-recall plots showing the performance of single-block (blue) *vs*. multi-block (red) calibrated models on simulated multi-OMICs data

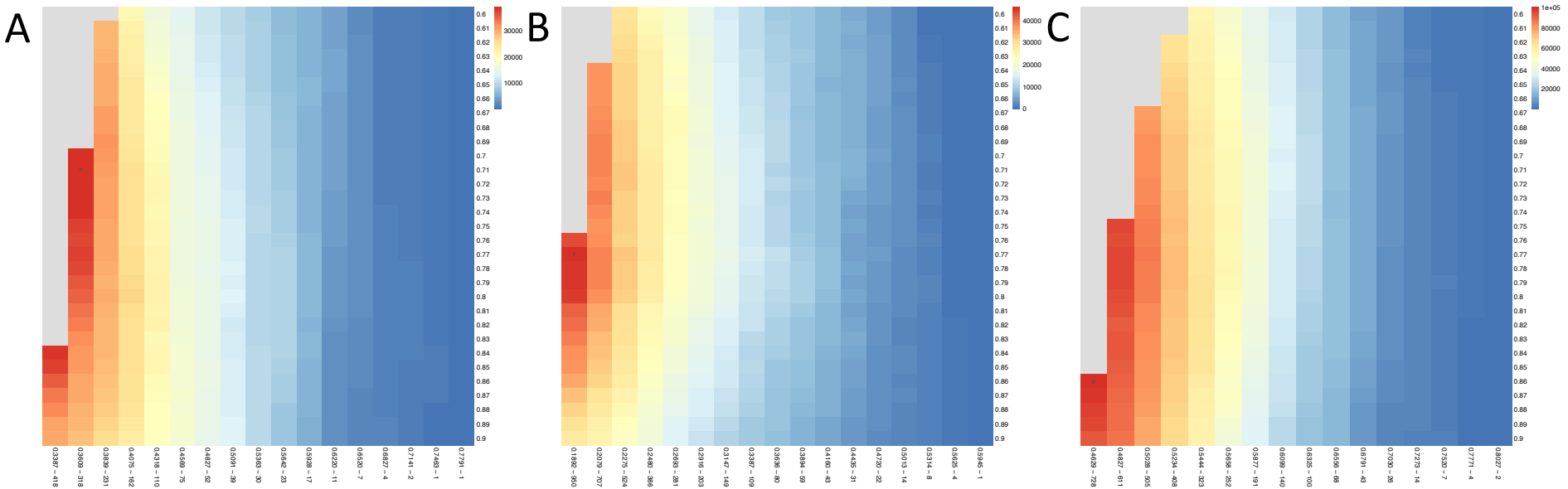


$\Rightarrow$ Clear increase in performance with multi-block calibration in all three blocks

$\Rightarrow$ Allowing for block-specific parameters improves performances on heterogeneous data

Integration with gene expression

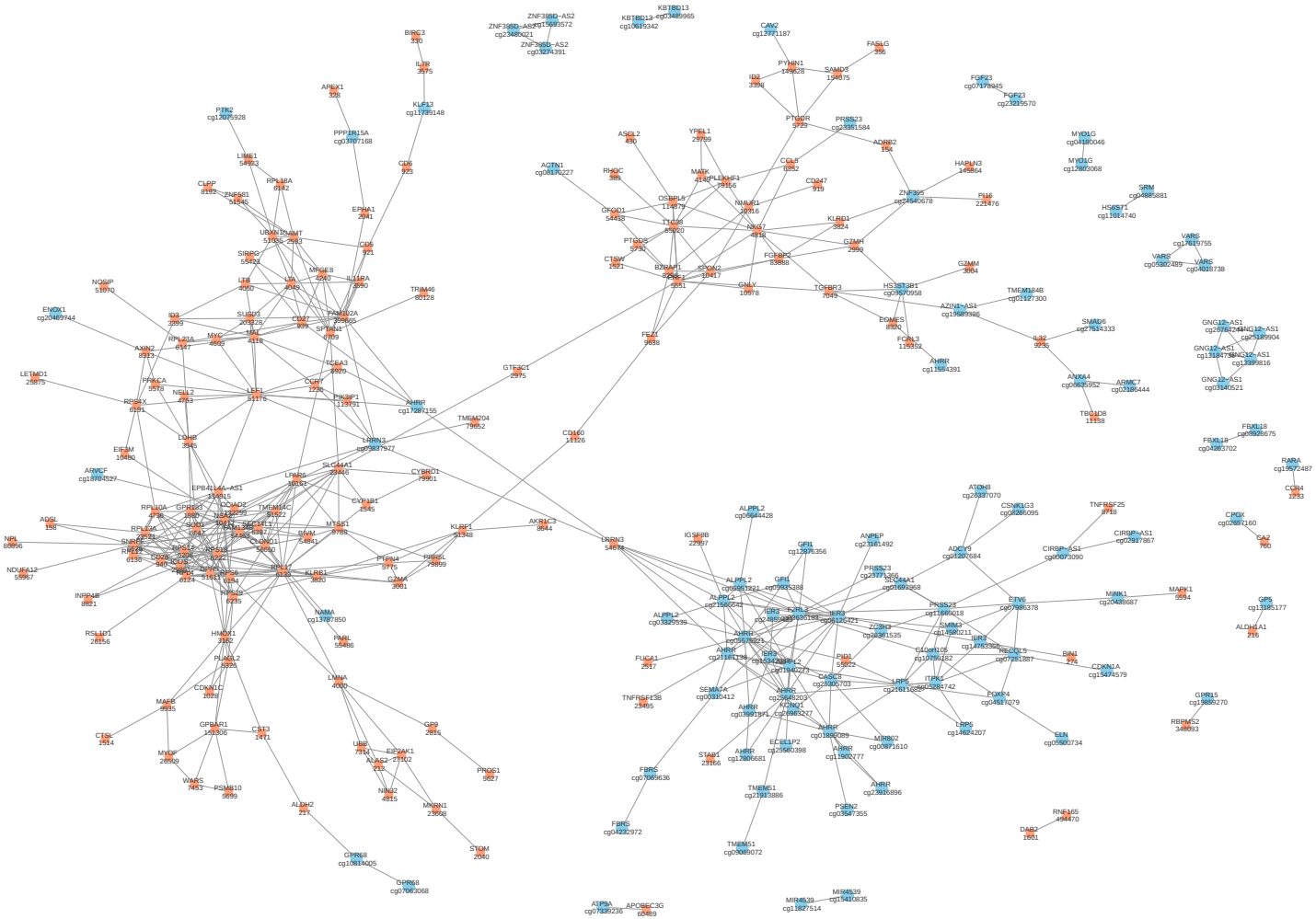
- Multi-block calibration procedure maximising block-specific stability scores



$\Rightarrow$ Calibration of the three pairs of parameters (λ_b, π_b)

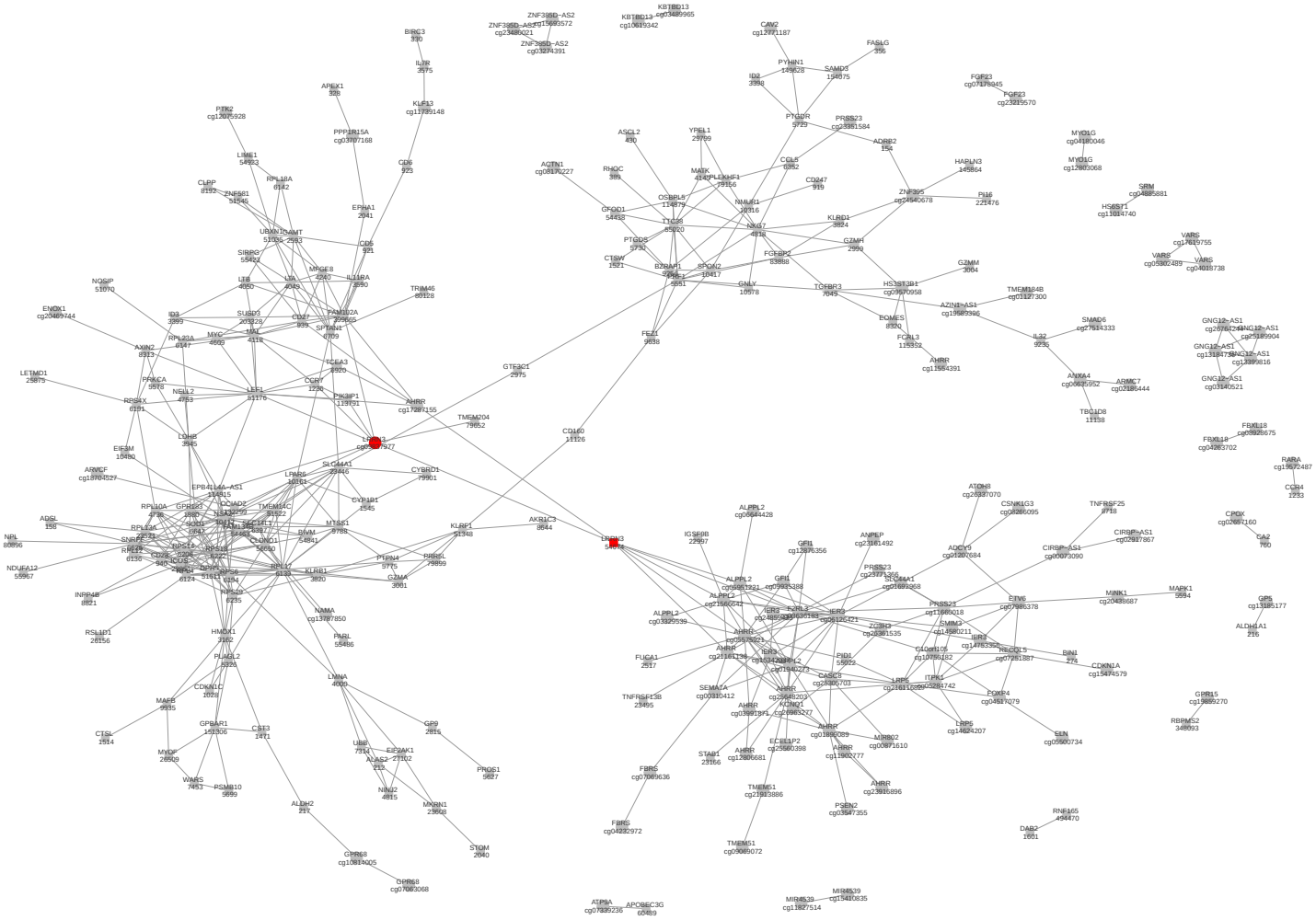
Multi-OMICs network: methylation and gene expression

- Integration of the 159 DNA methylation (blue) and 208 gene expression (red) markers of tobacco smoking measured in the same 250 individuals (NOWAC)



$\Rightarrow$ Detection of 96 cross-OMICs edges

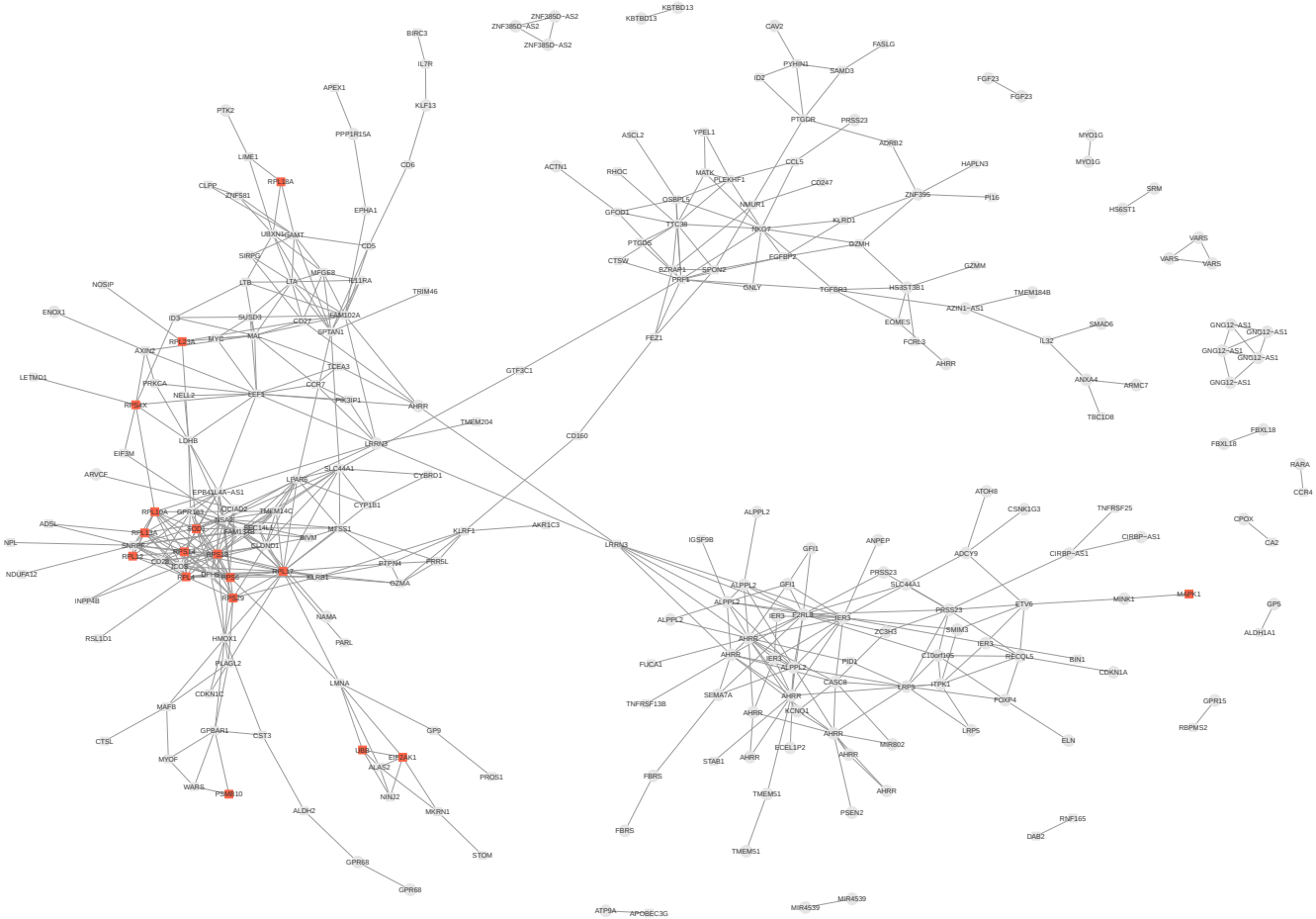
Central role of LRRN3



⇒ Very central role of LRRN3 (CpG site and transcript): linking the two blocks

Functional annotation

- Biological annotation of the transcripts using the Reactome knowledgebase



$\Rightarrow$ Same cluster of transcripts involved in the cellular response to stress

$\Rightarrow$ Annotation of related transcripts can give insight into functional role of the CpG sites

Conclusions/Perspectives

- Stability selection: complementing variable/edge selection algorithms with resampling procedures
- Calibration of the model based on a stability score measuring how far the model is from uniform (uninformative) selection of the features
- Optionally: constraint on the expected number of falsely selected features
- Multi-block extension to accommodate heterogeneous data sources

- Enhanced performances of the models compared to non-stability approaches (simulation studies)
- Limited increase in computation time compared to CV procedure
- Generated results seem biologically meaningful (functional annotation of the network)

  $\Rightarrow$ More work for a generalisable module annotation tool in networks (WGCNA)

Applications

- R package in preparation
- Approach is readily applicable to any variable selection model

  $\Rightarrow$ Could be used in combination with sparse (group) PLS models

- Stability selection models currently used for:
  - Detection of metabolomic features associated with a change in cognitive score (Nina)
  - Detection of proteomic/transcriptomic features associated with eosinophilic/neutrophilic status in asthma (Khezia)
  - Characterisation of the metabolomic signature of the BHS (Ana)
  - Identification/ranking of risk factors for cardiovascular disease prediction (Matt, Josh)
  - Identification of age-specific COVID-19 symptoms (Matt, Josh)

Acknowledgements

- Prof Marc Chadeau-Hyam (PI)
- Dr Sarah Filippi
- Dr Julien Chiquet
- Dr Stéphane Robin
- Prof Torkjel Sandanger
- Dr Therese Haugdahl Nost

MRC Centre for Environment & Health — Medical Research Council — Imperial College London

CANCER RESEARCH UK

AgroParisTech — INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT — PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

INRA SCIENCE & IMPACT

- Funding:
  - Cancer Research - UK 'Mechanomics' PRC project grant (Grant PRC 22184 to MC-H)
  - PhD Studentship from the MRC Centre for Environment and Health